Choice Models in Operations

Lecture 11: EM algorithm and Frank-Wolfe algorithm

Instructor: Srikanth Jagabathula *Scribe: JinHyun Kim*

Last week: Derived the MM algorithm for estimating the parameters of the MNL model through maximizing the log-likelihood function.

Necessary & Sufficient condition for the existence of a unique & bounded optimal solution to

$$\max_{\underline{\beta}} \sum_{t=1}^{T} log \frac{e^{\underline{\beta}^T \underline{z}_{j_t,t}}}{\sum_{i\in S_t} e^{\underline{\beta}^T \underline{z}_{i,t}}}$$

is that $\underline{\gamma}^T(\underline{z}_{i,t} - \underline{z}_{j_t,t}) \leq 0 \quad \forall i \in S_t, j_t, t = 1,2,...,T \quad \implies \quad \underline{\gamma} = \underline{0}$.

proof: [Necessity] Suppose $\exists$ a $\underline{\gamma} \neq \underline{0}$ such that $\underline{\gamma}^T(\underline{z}_{i,t} - \underline{z}_{j_t,t}) \leq 0 \quad \forall i \in S_t, j_t, t$.
Consider

$$l_t(\beta) = log \frac{e^{\underline{\beta}^T \underline{z}_{j_t,t}}}{\sum_{i\in S_t} e^{\underline{\beta}^T \underline{z}_{i,t}}} = -log \sum_{i\in S_t} e^{\underline{\beta}^T(\underline{z}_{i,t} - \underline{z}_{j_t,t})}$$

Then, we have for any $c > 0$,

$$l_t(\underline{\beta} - c\underline{\gamma}) = log \sum_{i\in S_t} e^{\underline{\beta}^T(\underline{z}_{i,t}-\underline{z}_{j_t,t}) - c\underline{\gamma}^T(\underline{z}_{i,t}-\underline{z}_{j_t,t})} = -log(1 + \sum_{i\in S_t\setminus\{j_t\}} e^{\underline{\beta}^T(\underline{z}_{i,t}-\underline{z}_{j_t,t}) - c\underline{\gamma}^T(\underline{z}_{i,t}-\underline{z}_{j_t,t})})$$

We have 2 cases:
(i) $\exists$ an i, j_t such that $\underline{\gamma}^T(\underline{z}_{i,t} - \underline{z}_{j_t,t}) < 0$.
Let $g_t(c) = l_t(\underline{\beta} - c\underline{\gamma})$.

$$g_t'(c) = \frac{\sum_{i\in S_t\setminus\{j_t\}} \underline{\gamma}^T(\underline{z}_{i,t} - \underline{z}_{j_t,t}) e^{\underline{\beta}^T(\underline{z}_{i,t}-\underline{z}_{j_t,t}) - c\underline{\gamma}^T(\underline{z}_{i,t}-\underline{z}_{j_t,t})}}{1 + \sum_{i\in S_t\setminus\{j_t\}} e^{\underline{\beta}^T(\underline{z}_{i,t}-\underline{z}_{j_t,t}) - c\underline{\gamma}^T(\underline{z}_{i,t}-\underline{z}_{j_t,t})}}$$

Because the derivative w.r.t c is always < 0, decreasing the value of c will strictly increase the value of $l_t(\underline{\beta} - c\underline{\gamma}) \quad \implies \quad$ the optimal solution cannot be bounded.
(ii) Suppose $\underline{\gamma}^T(\underline{z}_{i,t} - \underline{z}_{j_t,t}) = 0 \quad \forall i \in S_t, j_t, t = 1,2,...,T$.

$$\implies \quad l(\underline{\beta} - c\underline{\gamma}) = \sum_t l_t(\underline{\beta} - c\underline{\gamma}) = \sum_t l_t(\underline{\beta}) = l(\underline{\beta}) \quad \forall c \in \mathbb{R}$$

$$\implies \quad multiple\ optima$$

EM algorithm as a special case of MM algorithm

EM setup: Suppose we observe a data point $\underline{y}$ according to some distribution parametrized by $\underline{\theta}$.

Goal: Estimate $\underline{\theta}$ via MLE, i.e. solve $\underset{\theta}{max}\ log\ p(\underline{y}|\underline{\theta})$ where $p(\cdot|\underline{\theta})$ is the distribution according to which $\underline{y}$ was generated.
Suppose there is a latent variable such that if the variable is observed, then the MLE problem becomes simple(r). Let $\underline{z}$ denote the latent variable and let $p(\underline{y}, \underline{z}|\underline{\theta})$ denote the joint probability distribution function.
Then, the complete-data log-likelihood function is defined as

$$l_c(\underline{y}, \underline{z}) = log\ p(\underline{y}, \underline{z}|\underline{\theta})$$

Correspondingly, we refer to $log\ p(\underline{y}|\underline{\theta})$ as the incomplete-data log-likelihood funtion, denoted by $l_{IC}(\underline{y})$.

$$\underset{\underline{\theta}}{max}\ l_{IC}(\underline{y}; \underline{\theta}) = \underset{\underline{\theta}}{max}\ log\ p(\underline{y}|\underline{\theta}) = \underset{\underline{\theta}}{max}\ log \sum_{\underline{z}} p(\underline{y}, \underline{z}|\underline{\theta})$$

Let's apply the MM meta algorithm.
Let $\underline{\theta}^{(t)}$ be the current iterate.
The, we want to find $g(\cdot|\underline{\theta}^{(t)})$ that is a minorizing function, i.e. $l_{IC}(\underline{y}; \underline{\theta}) \geq g(\underline{\theta}|\underline{\theta}^{(t)})$ and $l_{IC}(\underline{y}; \underline{\theta}^{(t)}) = g(\underline{\theta}^{(t)}|\underline{\theta}^{(t)})$.

Trick: Suppose we have a function $f(\cdot)$ that is strict concave and say our goal is to find a minorizing function for $f(\sum_i x_i)$ at the current iterate $\underline{x}^{(t)}$. Since $f(\cdot)$ is concave, we must have $f(\sum_i \alpha_i y_i) \leq \sum_i \alpha_i f(y_i) \quad \forall \alpha_i \geq 0\ \forall i,\ \sum \alpha_i = 1$. Equality holds iff $y_i = y_j \quad \forall i,\ j$.
Set $\alpha_i = \frac{x_i^{(t)}}{\sum_j x_j^{(t)}}$ and $y_i = \frac{x_i}{x_i^{(t)}} \sum_j x_j^{(t)}$.
Suppose $x_i^{(t)} > 0 \ \ \forall i$ and $\sum_j x_j^{(t)} > 0$

$$\implies \quad f(\sum_i x_i) = f(\sum_i \alpha_i y_i) \geq \sum_i \alpha_i f(y_i) = \sum_i \frac{x_i^{(t)}}{\sum_j x_j^{(t)}} f(x_i(\frac{\sum_j x_j^{(t)}}{x_i^{(t)}}))$$

equality occuring iff $\frac{x_i}{x_i^{(t)}} = \frac{x_j}{x_j^{(t)}} \ \ \forall i,\ t$.
Therefore, $g(\underline{x}|\underline{x}^{(t)}) \triangleq \sum_i x_i^{(t)} f(x_i \frac{\sum_j x_j^{(t)}}{x_i^{(t)}})$ is a minorizing function to $f(\sum_i x_i)$ at $\underline{x}^{(t)}$.
Now, we apply this general idea to our setting by taking $f(\cdot)$ to be $log(\cdot)$ and $x_{\underline{z}} = p(\underline{y}, \underline{z}|\underline{\theta}) \ \ \forall \underline{z}$.

$$\therefore \quad log(\sum_{\underline{z}} p(\underline{y}, \underline{z}|\underline{\theta})) \geq \sum_{\underline{z}} \frac{p(\underline{z}, \underline{y}|\underline{\theta}^{(t)})}{\sum_{\underline{z}'} p(\underline{z}', \underline{y}|\underline{\theta}^{(t)})} log[p(\underline{y}, \underline{z}|\underline{\theta}) \frac{\sum_{\underline{z}'} p(\underline{z}', \underline{y}|\underline{\theta}^{(t)})}{p(\underline{z}, \underline{y}|\underline{\theta}^{(t)})}]$$

$$= \sum_{\underline{z}} \frac{p(\underline{z}, \underline{y}|\underline{\theta}^{(t)})}{\sum_{\underline{z}'} p(\underline{z}', \underline{y}|\underline{\theta}^{(t)})} log\ p(\underline{y}, \underline{z}|\underline{\theta}) + constant = \sum_{\underline{z}} \frac{p(\underline{z}, \underline{y}|\underline{\theta}^{(t)})}{p(\underline{y}|\underline{\theta}^{(t)})}\ log\ p(\underline{z}, \underline{y}|\underline{\theta}) + constant$$

$$= \sum_{\underline{z}} p(\underline{z}|\underline{y}, \underline{\theta}^{(t)})\ log\ p(\underline{z}, \underline{y}|\underline{\theta}) + constant = \mathbb{E}_{\underline{Z}|\underline{y}, \underline{\theta}^{(t)}}[log\ p(\underline{Z}, \underline{y}|\underline{\theta})]$$

$\longrightarrow$ E-step (expectation-step)

We then maximize the expectation above:

$$\underline{\theta}^{(t+1)} = \underset{\underline{\theta}}{argmax} \mathbb{E}_{\underline{Z}|\underline{y},\underline{\theta}^{(t)}}[log\ p(\underline{Z},\underline{y}|\underline{\theta})]$$

$$\longrightarrow \text{M-step (Max-step)}$$

We use the EM framework to derive the estimation algorithm for a latent-class MNL model with K classes.

<u>Data</u>: We observe choice from m customers. For customer i, we observe choices $(j_{i,t}, S_{i,t})$ for $t \in T_i$. The log-likelihood function is

$$\max_{\underline{\alpha},\ \underline{\beta}:\ \sum \alpha_k = 1,\ \alpha_k \geq 0\ \forall k} \sum_i log\ [\sum_{k=1}^{K} \alpha_k \prod_{t \in T_i} \frac{e^{\beta_{k,j_{i,t}}}}{\sum_{l \in S_{i,t}} e^{\beta_{k,l}}}]$$

We introduce latent variable z_i which is the class membership of each customer i. The complete-data log-likelihood function can be written as

$$l_c(\underline{z},\underline{y};\underline{\theta}) = \sum_i \sum_k \mathbb{1}[z_i = k]\ log\ [(\prod_{t \in T_i} \frac{e^{\beta_{k,j_{i,t}}}}{\sum_{l \in S_{i,t}} e^{\beta_{k,l}}})\alpha_k]$$

$$= \sum_i \sum_k \mathbb{1}[z_i = k]\ [log\ \alpha_k + \sum_{t \in T_i} log\ \frac{e^{\beta_{k,j_{i,t}}}}{\sum_{l \in S_{i,t}} e^{\beta_{k,l}}}]$$

Current iterate $\underline{\alpha}^{(t)},\ \underline{\beta}^{(t)}$.

<u>E-step</u>:

$$\mathbb{E}_{\underline{z}|\underline{y},\underline{\theta}^{(t)}}[l_c(\underline{z},\underline{y};\underline{\theta})] = \mathbb{E}_{\underline{z}|Data,\underline{\alpha}^{(t)},\underline{\beta}^{(t)}}[\sum_i \sum_k \mathbb{1}[z_i = k]\ [log\ \alpha_k + \sum_{t \in T_i} log\ \frac{e^{\beta_{k,j_{i,t}}}}{\sum_{l \in S_{i,t}} e^{\beta_{k,l}}}]]$$

$$= \sum_i \sum_{k=1}^{K} h_{ik}^{(t)}[log\ \alpha_k + \sum_{t \in T_i} log\ \frac{e^{\beta_{k,j_{i,t}}}}{\sum_{l \in S_{i,t}} e^{\beta_{k,l}}}]$$

where $h_{ik}^{(t)} = \mathbb{E}_{\underline{z}|Data,\underline{\alpha}^{(t)},\underline{\beta}^{(t)}}[\mathbb{1}[z_i = k]] = Pr(z_i = k|Data, \underline{\alpha}^{(t)}, \underline{\beta}^{(t)})$

$$= \frac{Pr(Data_i|z_i = k, \underline{\alpha}^{(t)}, \underline{\beta}^{(t)})Pr(z_i = k)}{\sum_{k'} Pr(Data_i|z_i = k', \underline{\alpha}^{(t)}, \underline{\beta}^{(t)})Pr(z_i = k')}$$

$$= \frac{\alpha_k^{(t)} \prod_{t \in T_i}(e^{\beta^{(t)}_{k,j_{i,t}}} / \sum_{l \in S_{i,t}} e^{\beta^{(t)}_{k,l}})}{\sum_{k'} \alpha_{k'}^{(t)} \prod_{t \in T_i}(e^{\beta^{(t)}_{k',j_{i,t}}} / \sum_{l \in S_{i,t}} e^{\beta^{(t)}_{k',l}})}$$

<u>M-step</u>:

$$\max_{\underline{\alpha},\ \underline{\beta}:\ \sum \alpha_k = 1,\ \alpha_k \geq 0\ \forall k} \sum_i \sum_{k=1}^{K} h_{ik}^{(t)}[\ log\ \alpha_k + \sum_{t \in T_i} log \frac{e^{\beta_{k,j_{i,t}}}}{\sum_{l \in S_{i,t}} e^{\beta_{k,l}}}]$$

The above optimization problem is separable in $\underline{\alpha}$ & $\underline{\beta}$. Optimizing over $\underline{\alpha}$ we get

$$\alpha_k^{(t+1)} = \frac{\sum_i h_{ik}^{(t)}}{\sum_{k'} \sum_i h_{ik'}^{(t)}}$$

$$\beta_k^{(t+1)} = \underset{k}{argmax} \sum_i h_{ik}^{(t)} \sum_{t \in T_i} log \frac{e^{\beta_{k,j_{i,t}}}}{\sum_{l \in S_{i,t}} e^{\beta_{k,l}}} \qquad \forall k$$

<u>Implementation note</u>: In order to determine $\beta_k^{(t+1)}$, you can use $\beta_k^{(t)}$ as the initial solution. So, we can do the updates in a "lazy" fashion by running only <u>one</u> MM update step. We can write

$$\beta_{k,j}^{(t+1)} = \beta_{k,j}^{(t)} + log \frac{\sum_i h_{ik}^{(t)} \sum_{t \in T_i} \mathbb{1}[j = j_{i,t}]}{\sum_i h_{ik}^{(t)} \sum_{t \in T_i} \mathbb{P}_k^{(t)}[j|S_{i,t}]}$$

Frank-Wolfe algorithm for estimating a rank-based choice model

<u>set up</u>: We have n items. We have observation of the form

$f_{j,S}$ = fraction of purchases of item j when S was offered for a collection of offer sets $S_1, S_2, ..., S_m$.

<u>Model</u>: We assume that the data are generated as follows. The population is described by a generation distribution (PMF). Over all possible rankings/pref lists of the n items. In particular, λ_σ is the probability of sampling σ, where $\sum_\sigma \lambda_\sigma = 1,\ \lambda_\sigma \geq 0\ \ \forall \sigma$, when given an offer set S, the customer samples a preference list σ according to $\underline{\lambda}$ and chooses the most preferred item from S according to σ.

<u>Estimation</u>: We estimate the model through MLE.

$$l(\lambda) = \sum_{i=1}^{m} \sum_{j \in S_i} (log\ \mathbb{P}_\lambda(j|S_i)\ f_{j,S_i})$$

$$= \sum_{i=1}^{m} \sum_{j \in S_i} (log\ (\sum_\sigma \lambda_\sigma \mathbb{1}[\sigma(j) < \sigma(k)\ \forall k \in S_i \backslash \{j\}])\ f_{j,S_i})$$

$\equiv j$ is most preferred among items in S_i under σ

$$= \sum_{i=1}^{m} \sum_{j \in S_i} ((log \sum_\sigma \lambda_\sigma \mathbb{1}[\sigma; j; S_i])\ f_{j,S_i})$$

The MLE problem now become

$$\max_{\underline{\lambda}} \sum_{i=1}^{m} \sum_{j \in S_i} f_{j,S_i}\ log(\sum_\sigma \lambda_\sigma \mathbb{1}[\sigma, j, S_i])$$

$$s.t.\ \sum_\sigma \lambda_\sigma = 1 \quad \lambda_\sigma \geq 0 \quad \forall \sigma$$

Because the objective is concave in the variable λ_σ and the constraints are linear, the above optimization problem is a convex program, albeit a large dimensional one.

Remarks:
1. In general, the above program has multiple optima. For tractability reasons, we choose a solution that has a small support size, where the support of $\underline{\lambda}$ is defined as $\{\sigma : \lambda_\sigma > 0\}$.
2. Consider the following reformulation of the optimization problem:

$$\max_{\underline{\lambda}, \underline{g}} \sum_{i=1}^{m} \sum_{j \in S_i} f_{j,S_i} \; log \; g_{j,S_i}$$

$$s.t. \; g_{j,S_i} = \sum_{\sigma} \lambda_\sigma \mathbb{1}[\sigma; j; S_i] \;\; \forall j \in S_i, \; i = 1, ..., m$$

$$\sum_{\sigma} \lambda_\sigma = 1, \; \lambda_\sigma \geq 0 \;\; \forall \sigma$$

Now consider the following vectorization. Let $L = \sum_{i=1}^{m} |S_i|$ and $\underline{g} \in \mathbb{R}^L$ s.t. $(\underline{g})_{j,S_i} = g_{j,S_i}$. Also, for any σ, let $\underline{e}_\sigma \in \{0,1\}^L$ s.t. $(\underline{e}_\sigma)_{j,S_i} = \mathbb{1}[\sigma; j; S_i]$.
We can rewrite the above optimization problem as

$$\max_{\underline{g}} \sum_{i=1}^{m} \sum_{j \in S_i} f_{j,S_i} \; log \; g_{j,S_i}$$

$$s.t. \; \underline{g} \in conv(\{\underline{e}_\sigma : \sigma\})$$

Remarks:
Suppose $\underline{g}^*$ is an optimal solution to the above program. It follows from Caratheodory's Theorem that $\exists$ a convex decomposition of $\underline{g}^*$ in terms of $\underline{e}_\sigma$ with support size at most $L + 1$.